美团云对象存储系统

尹殷
yinyin@meituan.com

美团云

目录

云计算与存储系统

- 云计算平台存储产品
  - 主机本地存储
  - 弹性块存储（EBS）
  - 对象存储（S3）
  - KV与数据库

	读写延迟	可用性	可扩展性	按量付费	多点读写	长度适配
主机本地存储	低	低	无	✘	✘	通用
弹性块存储（EBS）	低	高	中	✘	✘	通用
对象存储（S3）	高	高	高	✓	✓	>百KB
KV与数据库	中	高	高	✓	✓	<百KB

云计算与存储系统

- 云计算平台存储产品应用场景
  - 对象存储（S3）：
    - 内容存储和分发（图片、视频、网站静态资源）
    - 数据分析的存储
    - 备份、归档和灾难恢复
    - 静态网站托管
    - 主机镜像

云计算与存储系统

- 对象存储的特点
  - Key-Value：用户指定Key
  - 一次写入多次读取，少量更新
  - 小对象与大对象共存（万亿个，几十KB~几TB）
  - HTTP接口
  - AWS-S3事实标准
  - 账户-桶管理结构
  - 权限体系

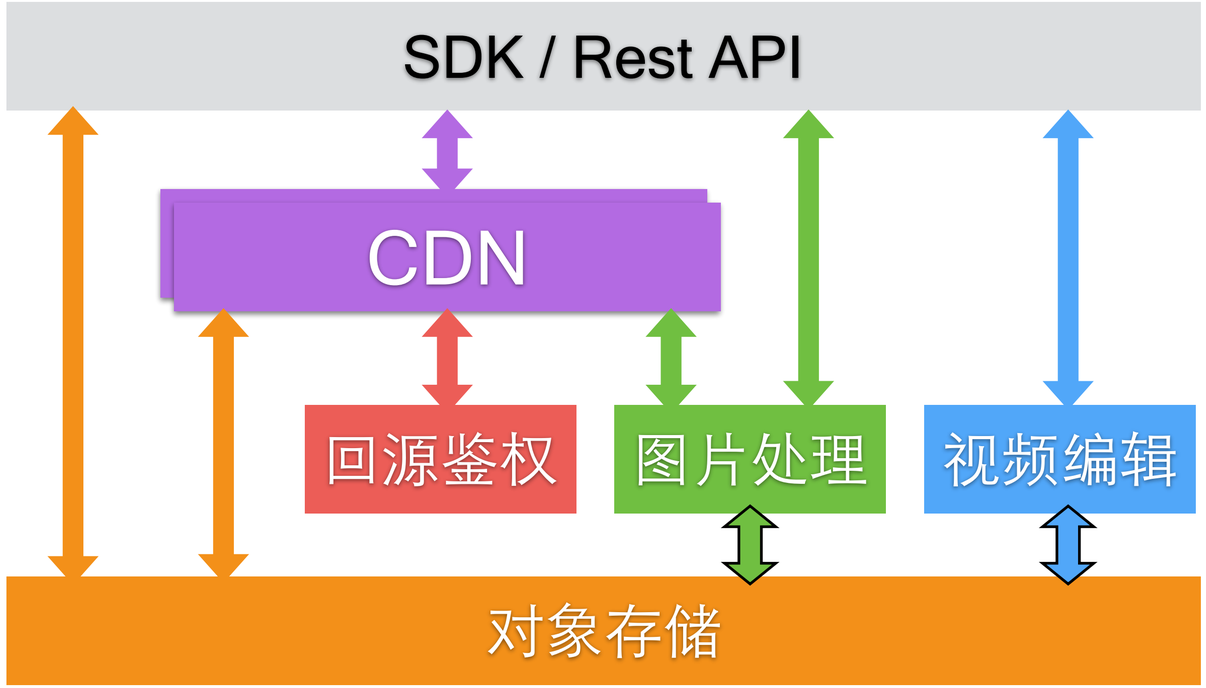
目录

美团云存储现状

- 集群规模
  - 100+台服务器，3PB总容量，dx+yf+cq三机房部署
  - 1.5PB存储量
  - 4.2亿对象
  - 日访问峰值过4亿
  - 服务近百个业务线：cos，大象，云盘，it，金融，外卖，旅游，猫眼，饭否
  - SDK：Java PHP Python Ruby Nodejs

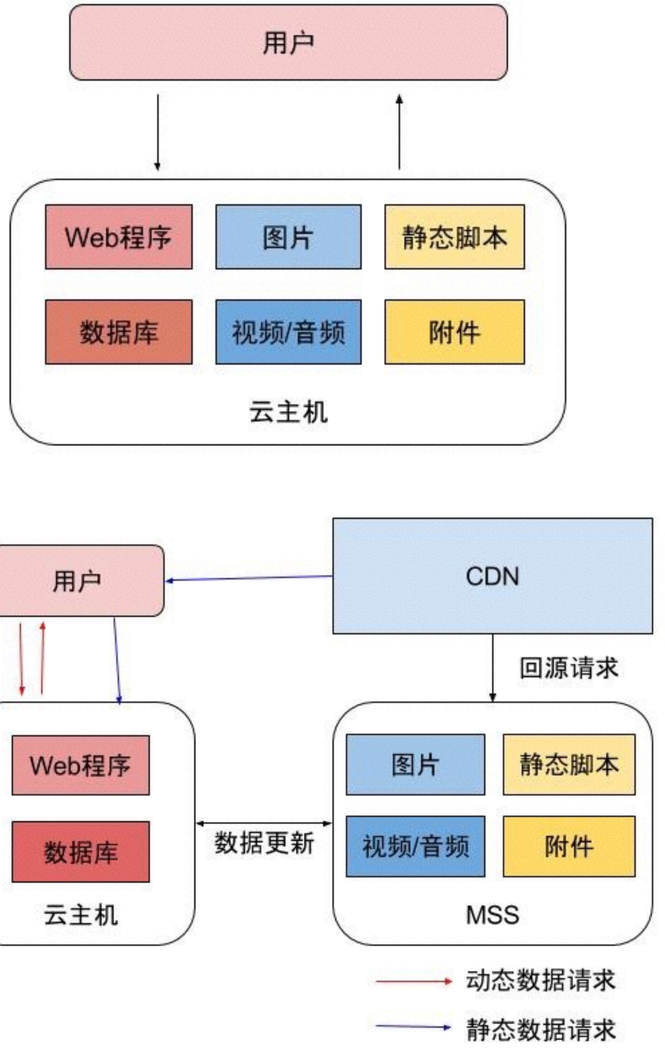
美团云存储现状

- 公有云产品线
  - 对象存储
  - CDN加速
  - 视频转码
  - 图片服务
  - 回源鉴权
  - 镜像源站



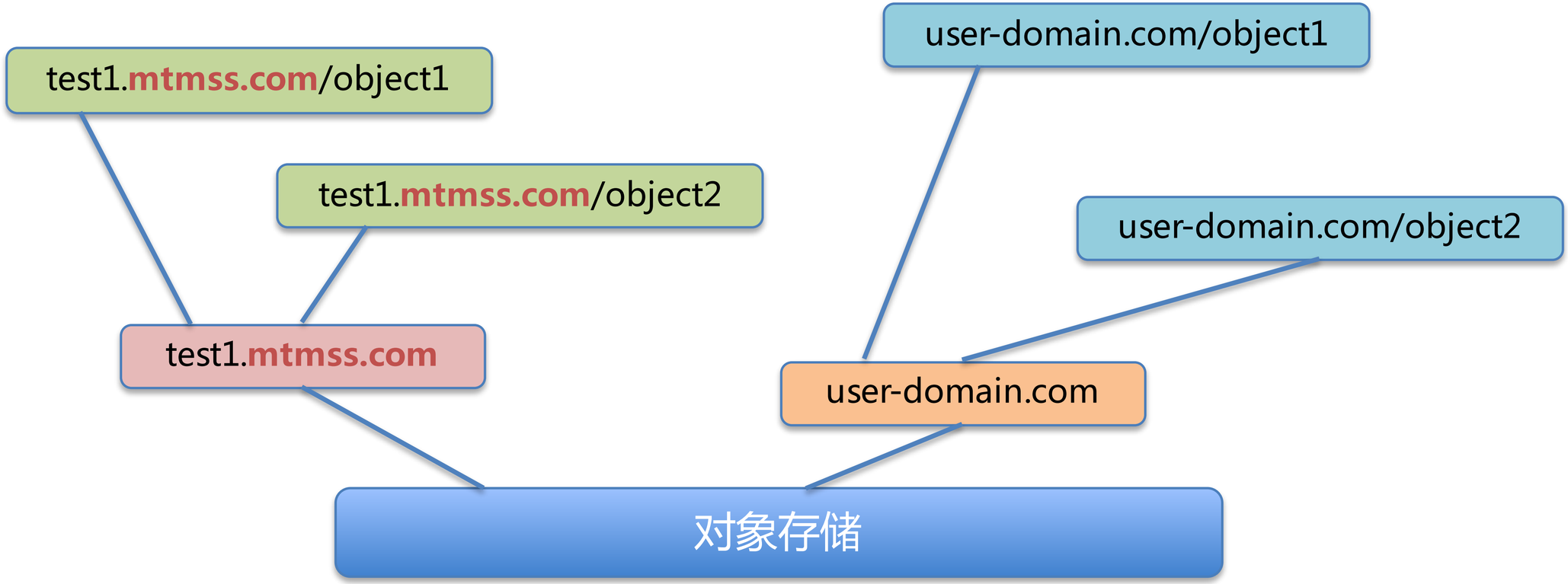
美团云存储现状

- 如何使用对象存储
  - 传统网站架构：动态数据、静态数据不分离
  - 使用对象存储、CDN
    - Web服务负载低
    - 海量存储空间
    - 存储费用低



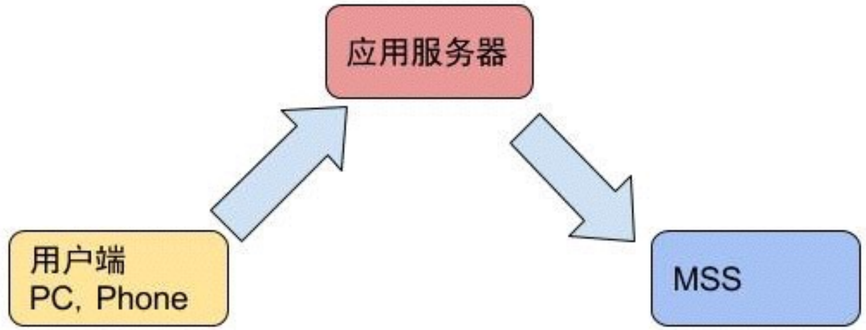
美团云存储现状

- S3存储模型



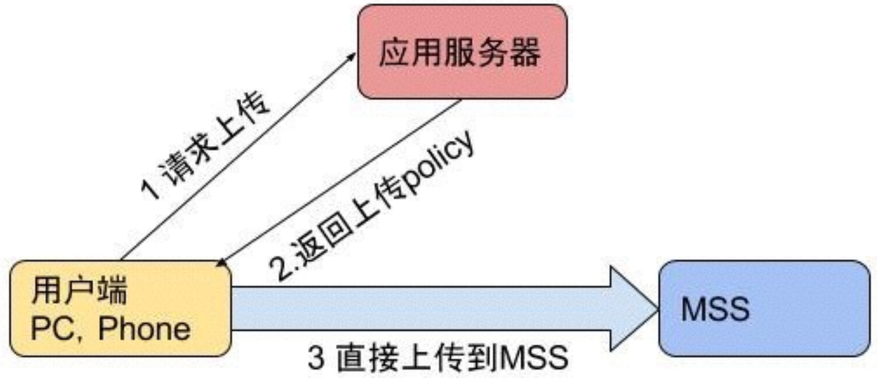
美团云存储现状

- MSS直传实践
  - 背景：通过业务服务器上传
  - 缺点：
    - 上传慢
    - 服务不稳定
    - 难以扩展



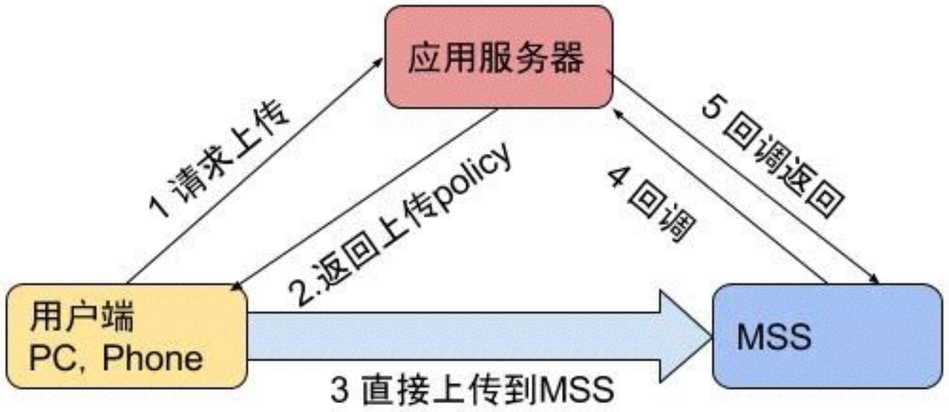
美团云存储现状

- MSS直传实践
  - 方案A:服务端签名
  - 支持PreSign或表单
  - 数据不经过应用服务器



美团云存储现状

- MSS直传实践
  - 方案B:服务端签名加回调
  - 支持表单
  - 应用服务器需要获取文件名，大小等；回调支持魔法变量

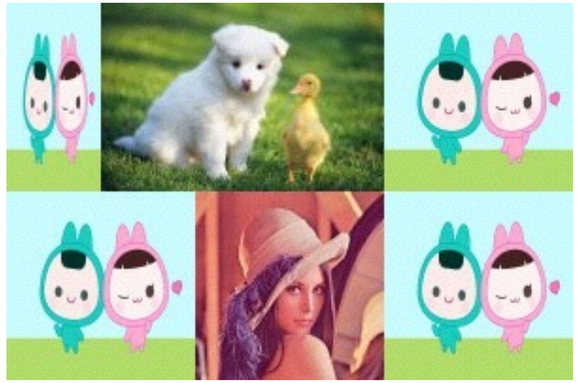


美团云存储现状

- 图片服务

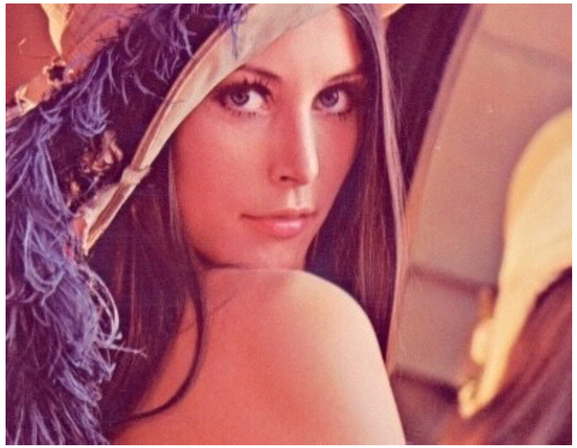
- 图片缩放、裁剪
- 图片旋转、自适应方向
- 质量变换、格式转换
- 拼图、水印，图片水印，文字水印，图文混合水印
- 管道、样式

拼图



质量变换@10q

裁剪@100-200-400-0a



缩放 @596_11.png

美团云存储现状

- 视频转码服务
- 视频格式转换
- m3u8切片，私有m3u8
- 视频采样缩略图
- 音视频元信息

```
#EXTM3U
#EXT-X-VERSION:3
#EXT-X-TARGETDURATION:17
#EXT-X-MEDIA-SEQUENCE:0
#EXTINF:12.705900,
http://msstest-corp.sankuai.com/privatemedia/tesppm3u8.m3u8/0.ts?
AWSAccessKeyId=93b6ab9e70284b32824a974c521155e9&Expires
=1458405012&Signature=n8hmq4oVmK50RR5GMc2d3f0%2FHow
%3D
#EXTINF:13.666667,
http://msstest-corp.sankuai.com/privatemedia/tesppm3u8.m3u8/1.ts?
AWSAccessKeyId=93b6ab9e70284b32824a974c521155e9&Expires
=1458405012&Signature=IjLFVMRMewo6bvW%2FEc9yeUXZq
%2F8%3D
#EXTINF:7.600000,
http://msstest-corp.sankuai.com/privatemedia/tesppm3u8.m3u8/2.ts?
AWSAccessKeyId=93b6ab9e70284b32824a974c521155e9&Expires
=1458405012&Signature=BMowcKg5xbNRewSXCxxv1iVIpAA%3D
#EXTINF:8.333333,
```

美团云存储现状

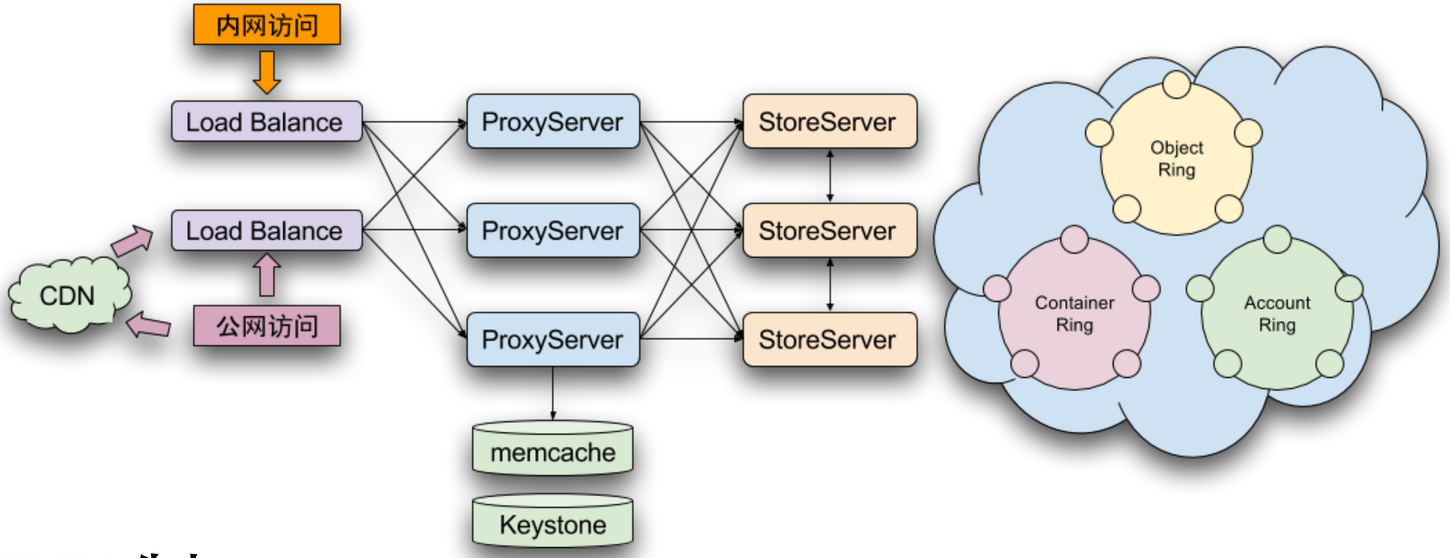
- 融合CDN
  - 刷新和预取
  - 回源鉴权
    - 私有bucket cdn加速
  - 日志分析
    - 请求统计，响应时间分布，连通率

美团云存储现状

- 内部典型业务场景
  - cos：敏感业务数据存储，业务鉴权+tempurl跳转
  - 云盘：海量对象存储，图片预览API，视频截图与编解码API，tempurl上传下载
  - MySQL备份：分片并发上传
  - 饭否、猫眼预告片：图床与CDN加速
  - 美团点评合并业务：镜像源站服务
  - Ops发布源：动态域名解析、内网限速
  - 点评UGC图片：对接Venus，数亿级别对象存储
  - 酒店、团购：app分发

美团云存储现状

- 基于Openstack/Swift构建
  - 无中心化设计
  - 一致性Hash
  - 最终一致性模型
  - 没有将对象聚合存储
  - 单个Bucket对象总数量限制
  - 纠删码性能
  - Python项目性能和可维护性



目录

下一代对象存储系统设计

- Swift的问题
  - ✤使用一致性哈希算法分布数据，最终一致性
  - ✤没有聚合文件，浪费inode, EC效率低
  - ✤S3协议难以完全兼容
- 为何要自研对象存储系统
  - ✤云存储是云计算的基石（亚马逊S3系统）
  - ✤S3协议在业界使用广泛、生态完善
  - ✤现有线上系统使用的开源系统Swift有诸多问题

	扩容	数据修复速度	数据检查	海量小文件	支持EC	读写性能	运维
swift	慢，代价高	慢	弱	困难	困难	低	复杂
自研系统	简单，快速	快	强	简单	简单	高	简单

下一代对象存储系统设计

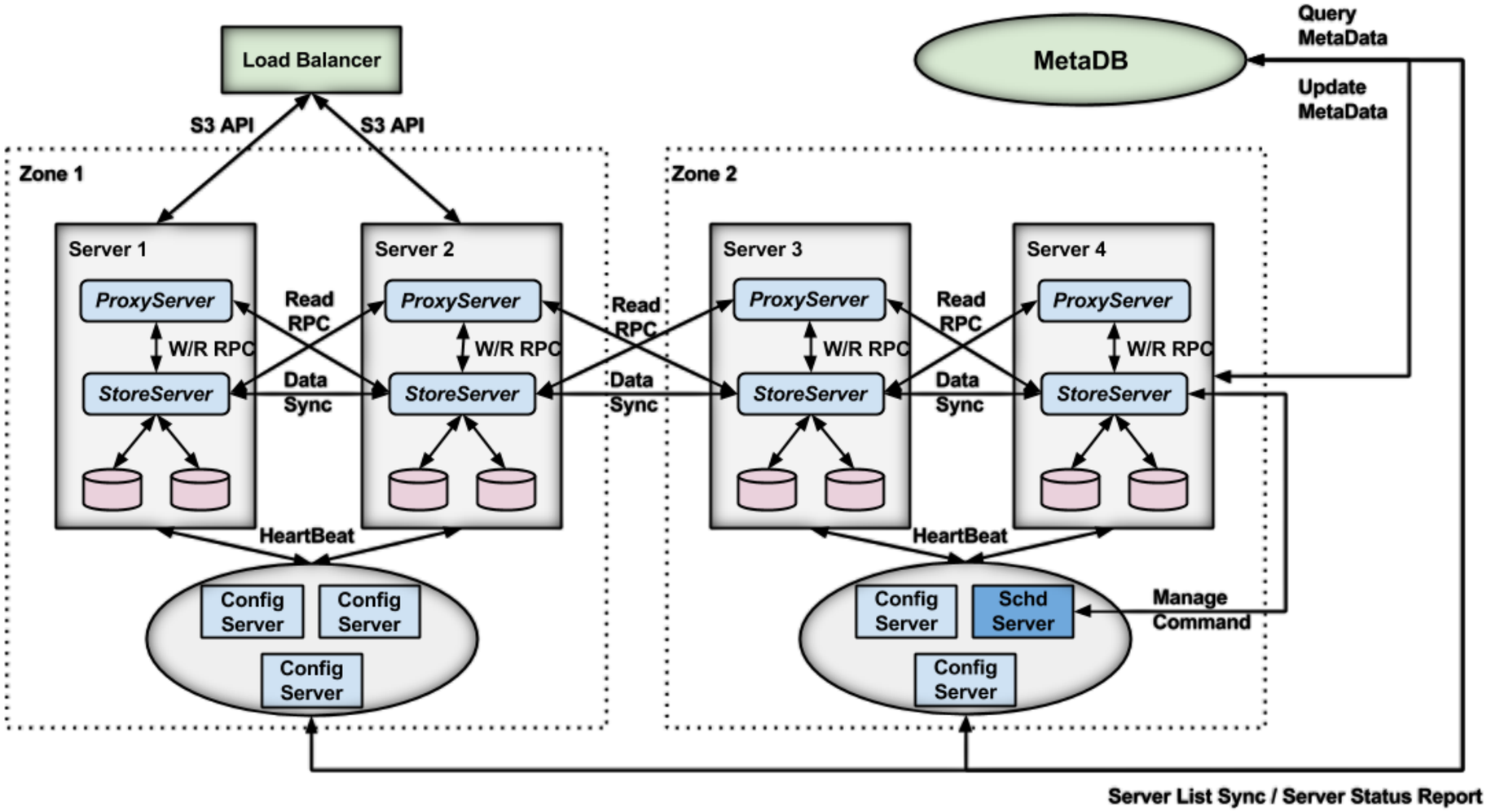
- 存储系统设计Key-Point
  - 数据分布与元数据
    - 运维、迁移恢复、元数据可扩展性
    - 一致性Hash：Swift，Ceph
    - 元数据集中存储：GFS，HDFS，Haystack
  - 高可用
    - Swift：计算在Hash Ring上的新位置
    - GFS：更换一组新的副本

下一代对象存储系统设计

- 存储系统设计Key-Point
  - 物理存储形式
    - 直接存储
    - 聚合存储
    - Update/Delete vs. Append Only
    - Replica vs. Erasure Code，EC恢复
  - 硬件环境
    - 万兆网卡
    - 低成本：低端CPU、高密度存储

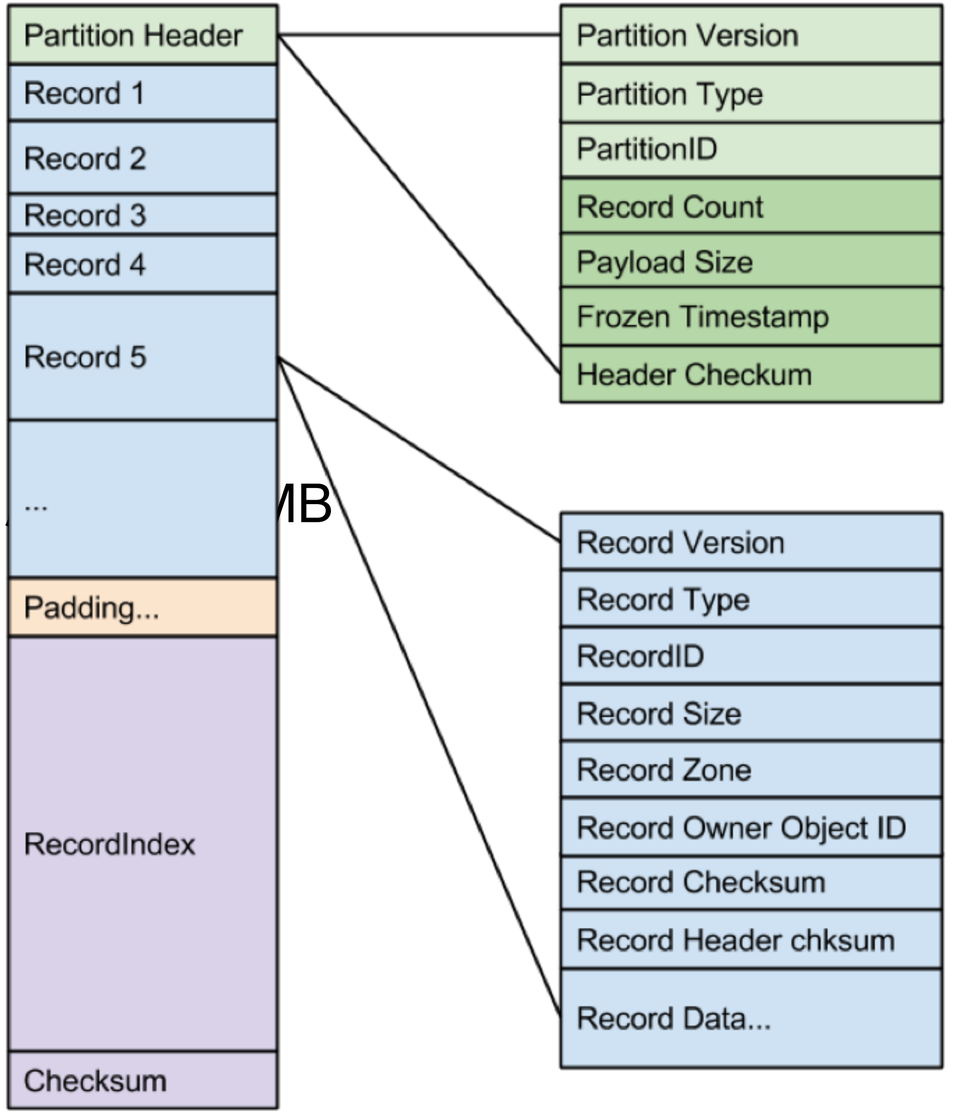
下一代对象存储系统设计

- 系统架构



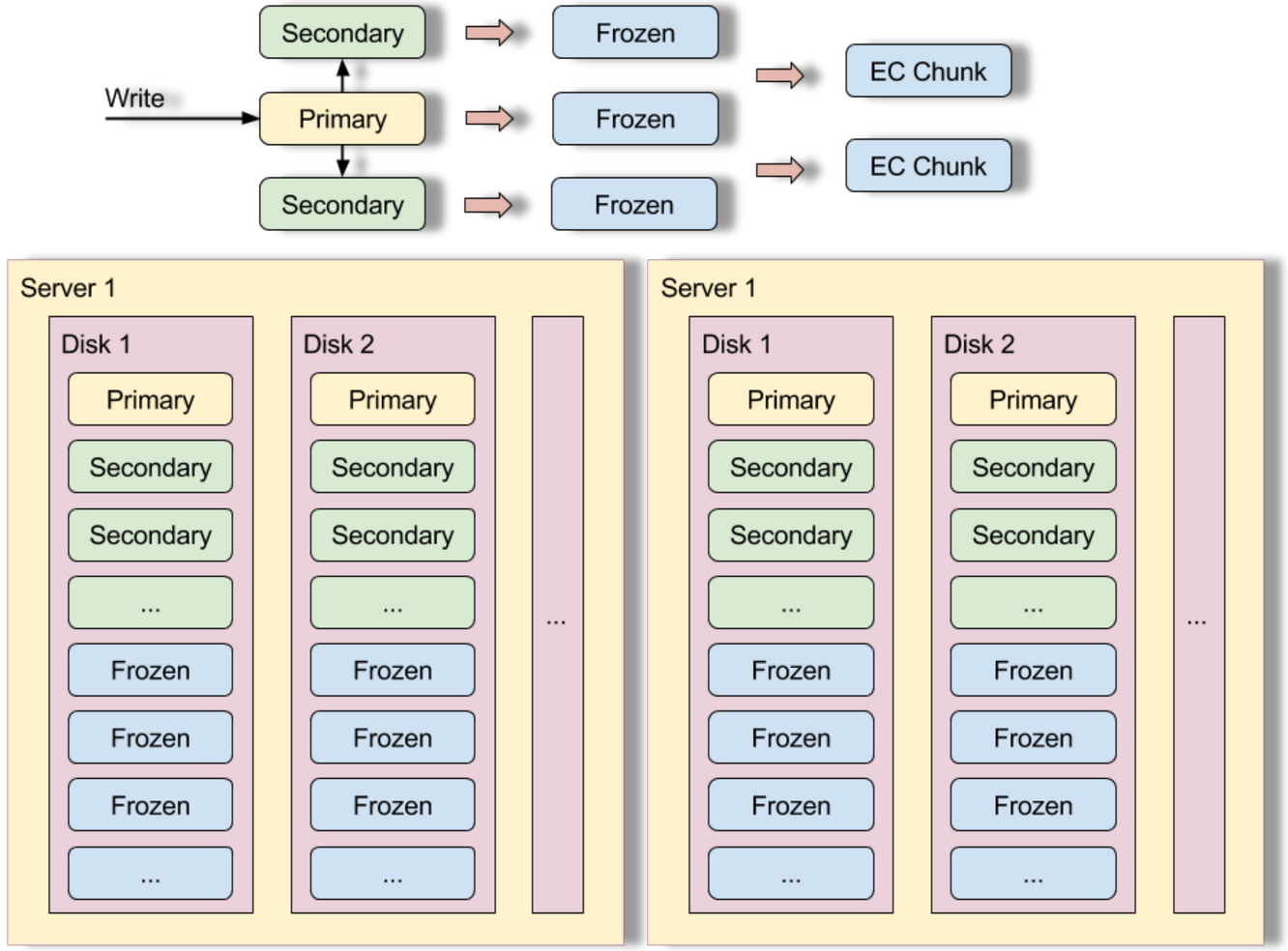
下一代对象存储系统设计

- 存储设计
  - Object
    - 拆分为Record
  - Partition
    - 负载平衡、迁移复制的最小单位；变长，最大长[illegible]MB
    - Primary / Secondary Partition
  - Record
    - 最大长度限制2MB；大文件拆分，流式写入
    - Group Commit
  - Record Index



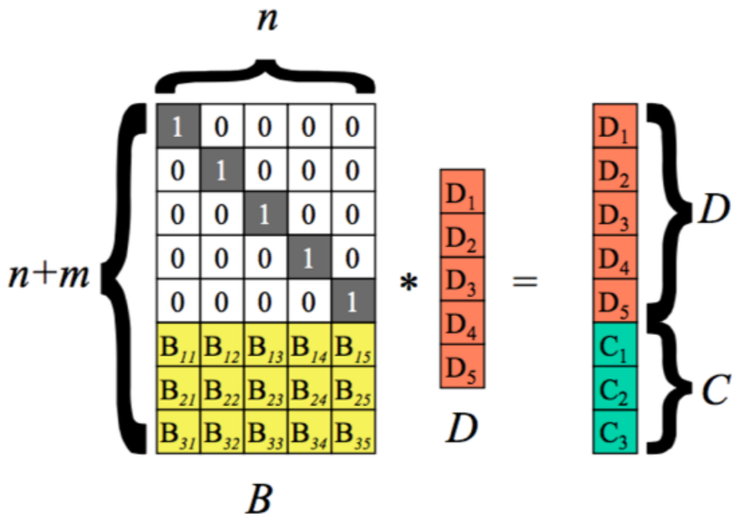
下一代对象存储系统设计

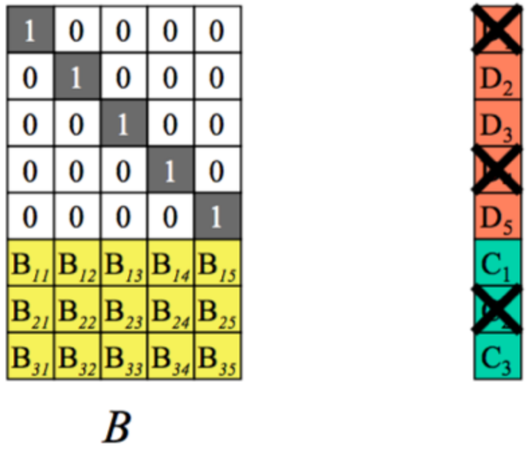
- Partition存储于状态转移



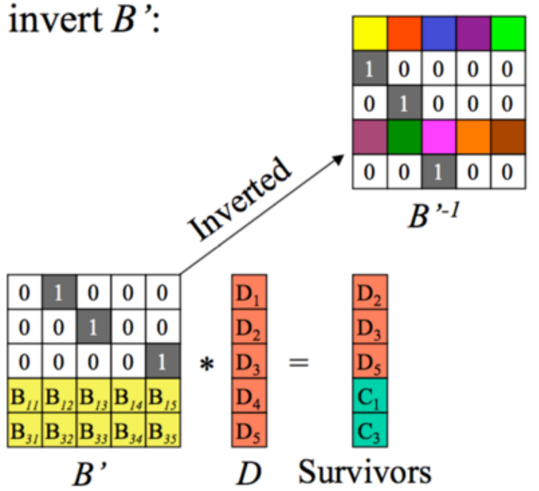
下一代对象存储系统设计

- RS纠删码
  - 使用生成矩阵计算和更新校验块
  - 使用高斯消元从故障中恢复数据
  - 使用Galois域进行代数计算



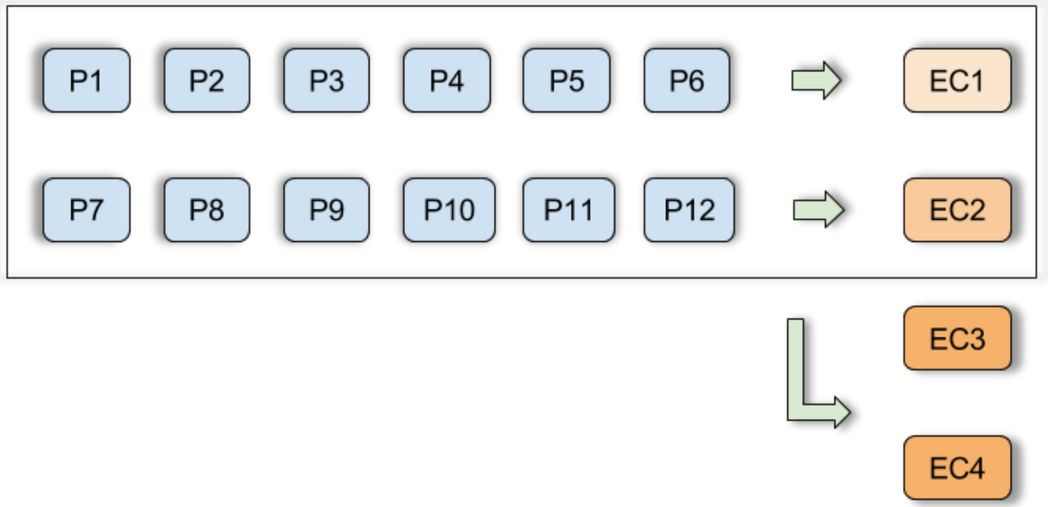


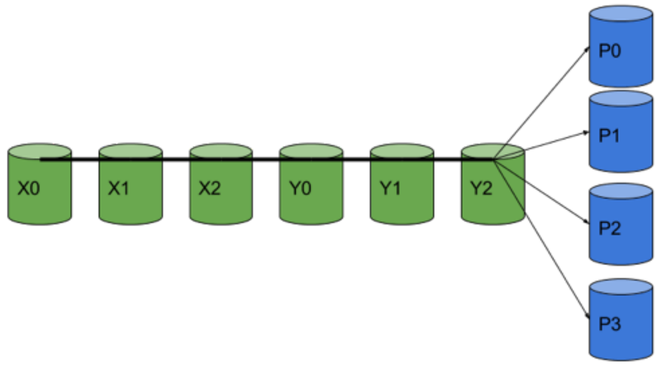
- Now, invert *B'*:



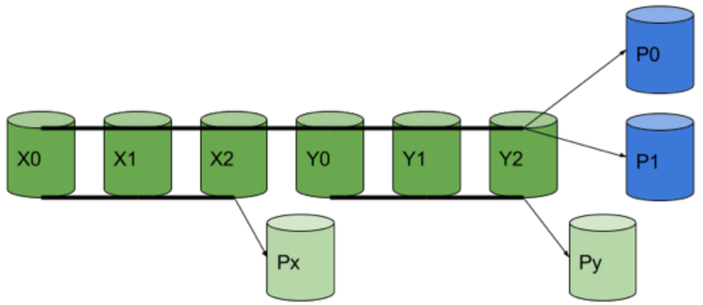
下一代对象存储系统设计

- LRC纠删码
  - 抛开恢复时间谈可靠性是耍流氓
  - EC实时恢复计算





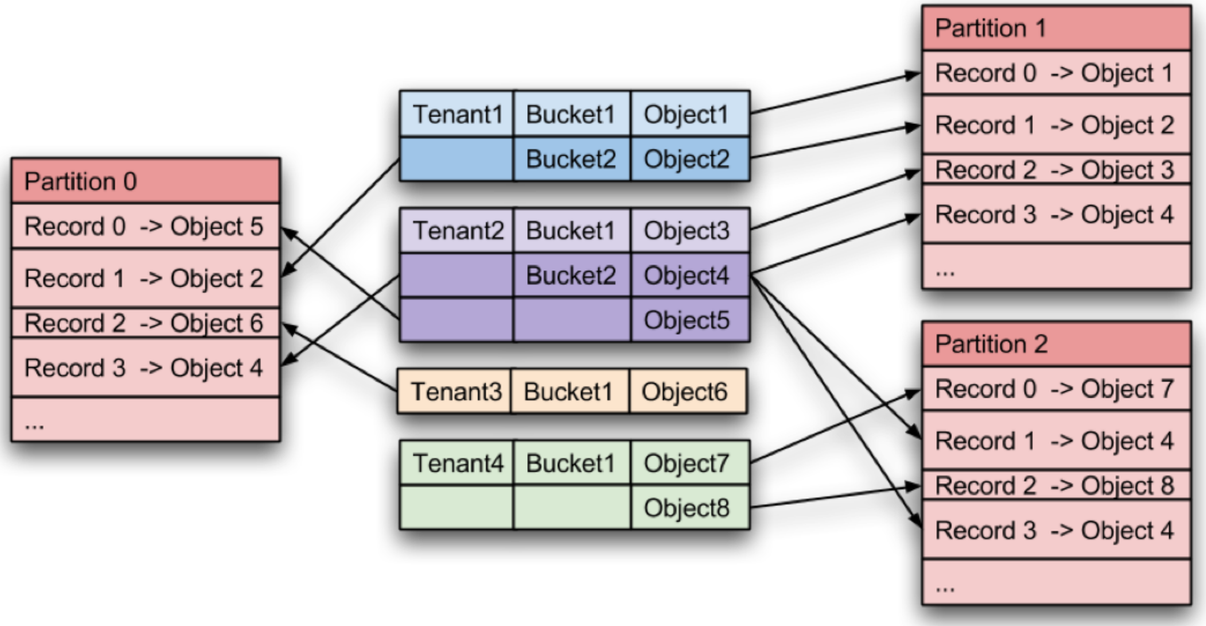
Reed-Solomon(6,4)

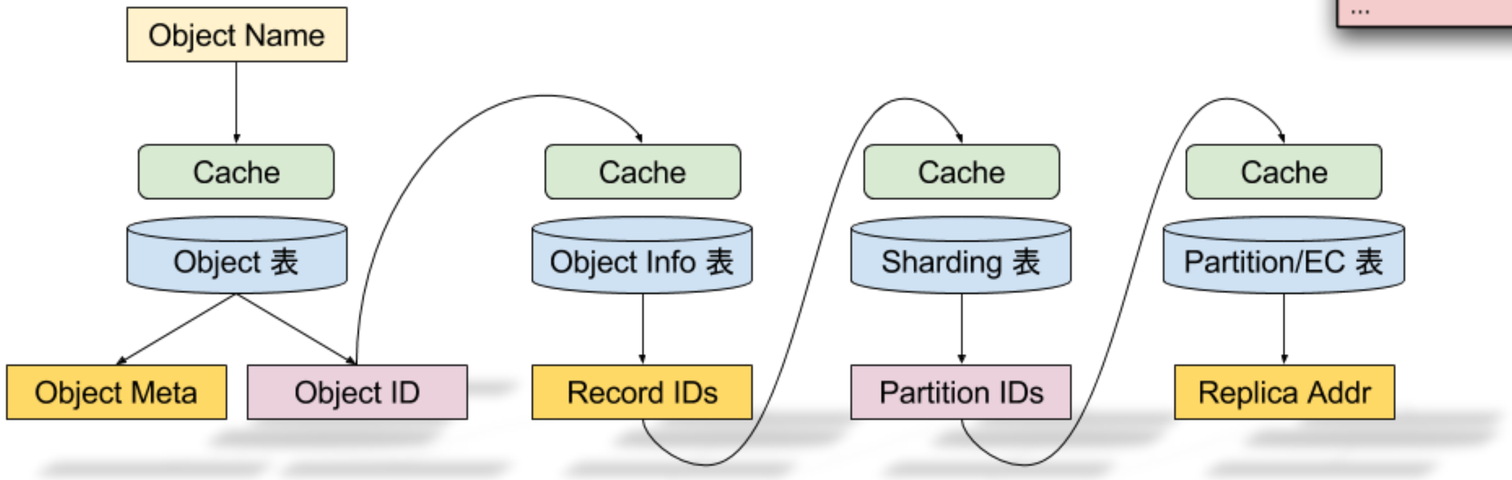


LRC(6,2,2)

下一代对象存储系统设计

- 元数据设计
  - Object Namespace
  - Object Meta数据
  - 对象覆盖写
  - 多版本存储
  - 数据去重
  - 缓存友好





下一代对象存储系统设计

- Sharding表设计
  - Record ID：[ServerIP + DiskID + Timestamp]
  - Erasure Code / Garbage Collection：sharding重组

Start Record ID	Partition ID	Type
90.12.40	90	Replica
...	...	...
100.9.80	100	Replica
100.10.1	101	Replica
100.10.20	102	Replica
100.10.50	103	Replica
100.10.80	104	Replica
100.11.30	105	Replica
...	...	...
101.0.50	180	Replica

Start Record ID	Partition ID	Type
90.12.40	90	Replica
...	...	...
100.9.80	100	Replica
100.10.1	1001	EC
100.10.40	1003	EC
100.10.80	104	Replica
100.11.30	105	Replica
...	...	...
101.0.50	180	Replica

下一代对象存储系统设计

- 技术展望
  - 更全面的AWS-S3接口兼容
  - 多Region与跨Region备份
  - 冷存储成本控制
  - GoLang并发网络框架

美团云
Meituan Open Services

专业提升效率 稳定创造价值

mos.meituan.com

美团云
Meituan Open Services